View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Research Papers in Economics

DEAKIN UNIVERSITY AUSTRALIA

Faculty of Business and Law
School of Accounting, Economics and Finance

School Working Paper

ECONOMICS SERIES

SWP 2009/13

Could It Be Better to Discard 90% of the Data? A Statistical Paradox

T.D. Stanley, Stephen B. Jarrell, and Hristos Doucouliagos

The working papers are a series of manuscripts in their draft form. Please do not quote without obtaining the author's consent as these works are in their draft form. The views expressed in this paper are those of the author and not necessarily endorsed by the School or *IBIS*World Pty Ltd.

Could It Be Better to Discard 90% of the Data? A Statistical Paradox

by

T.D. Stanley, Stephen B. Jarrell, and Hristos Doucouliagos

Abstract

Conventional practice is to draw inferences from all available data and research results, even though there is ample evidence to suggest that empirical literatures suffer from publication selection bias. When a scientific literature is plagued by such bias, a simple discarding of the vast majority of empirical results can actually improve statistical inference and estimation. Simulations demonstrate that, if the majority of researchers, reviewers, and editors use statistical significance as a criterion for reporting or publishing an estimate, discarding 90% of the published findings greatly reduces publication selection bias and is often more efficient than conventional summary statistics. Improving statistical estimation and inference through removing so much data goes against statistical theory and practice; hence, it is paradoxical. We investigate a very simple method to reduce the effects of publication bias and to improve the efficiency of summary estimates of accumulated empirical research results that averages the most precise ten percent of the reported estimates (*i.e.*, '*Top10*'). In the process, the critical importance of precision (the inverse of an estimate's standard error) as a measure of a study's quality is brought to light. Reviewers and journal editors should use precision as one objective measure of a study's quality.

Key Words: Publication Selection, Meta-analysis, Precision, Simulations, Meta-Regression.

Stanley: Department of Economics, Hendrix College, 1600 Washington St., Conway, AR, 72032, USA. Email: Stanley@hendrix.edu. Phone: 501-450-1276.

Jarrell: Department of Global Management and Strategy, College of Business, Western Carolina University, Cullowhee, NC 28723. Email: jarrell@email.wcu.edu. Phone: 828-631-1420.

Doucouliagos: School of Accounting, Economics and Finance, Deakin University, 221 Burwood Highway, Burwood, 3125, Victoria, Australia. Email: douc@deakin.edu.au. Phone: 61 03 9244 6531.

1. Introduction

Fundamental statistical criteria such as efficiency and power are fueled by data. To statisticians, data are sacrosanct. 'The more data the better' is the statistician's motto. Only under extreme conditions, termed 'outliers,' can a researcher be justified in ignoring or discarding the data. Even then, the removed data must be suspected to be in error: miscoded, mistakenly measured, or somehow gathered from an entirely different population. If such a principle of inclusion applies to data, in general, would it not also be applicable to the contents of the studies published in our most prestigious scientific journals?

This paper identifies a common condition in social science and medical research where it may be better to discard ninety percent of the reported empirical estimates, routinely. When is it better to discard 90% of the data? If the majority of the researchers, reviewers, and editors use statistical significance as a criterion for reporting or publishing an estimate, statistical inference and estimation may improve by completely ignoring the vast majority of the reported research. When reported estimates are selected for their statistical significance, valid empirical inference is threatened because the research base will contain large publication selection biases.

Publication selection bias has long been acknowledged as a severe threat to statistical inference and scientific practice (Sterling, 1959; Tullock, 1959; Feige, 1975; Rosenthal, 1979; Glass, McGaw and Smith, 1981; Lovell, 1983; Hedges and Oklin, 1985; Begg and Berlin, 1988; DeLong and Lang, 1992; Card and Krueger, 1995; Sterling, Rosenbaum and Weinkam, 1995; and Copas, 1999, to mention a few). When the majority of reported findings are selected for statistical significance, empirical phenomena can be manufactured, mere artifacts of the publication selection process. For example, the efficacy of a new pharmacological treatment or the adverse employment effect of raising the minimum wage may be seen by many researchers as established fact, when the effect is nothing more than the result of publication selection bias (Krakovsky, 2004; Doucouliagos and Stanley, 2009). In the social sciences, the overreliance and abuse of statistical significance has caused

a longstanding controversy and a revision of the American Psychological Association editorial policy (Carver 1978; Cohen, 1994; Harlow et al. 1997; Daniel, 1998; and APA, 1994). By examining 65 separate meta-analyses of separate areas of economics research, Stanley and Doucouliagos (2008) document how publication selection is a serious issue in two-thirds of empirical economics. Sterling, Rosenbaum and Weinkam (1995) show how selection for significance is also a widespread practice in the natural sciences. Gerber, Green and Nickerson (2001) and Gerber and Malhorta (2008) show the same in political science. A recent systematic review found evidence that publication selection is widespread in medical research (Hopewell *et al*., 2009). Due to the widely recognized adverse effects of publication selection bias, all the best medical journals now require the prior registration of clinical trials (Krakovsky, 2004).

The purpose of this paper is to document this statistical paradox that discarding 90% of the data might actually improve scientific inference. To reduce publication selection bias, we offer a very simple, 'back-of-an-envelope' remedy, the 'top 10 percent.' The performance of the average of the most precise ten percent of the reported estimates (*i.e.*, '*Top10*') is simulated and compared to alternative conventional summary statistics that use *all* the reported research results. In some realistic circumstances, the *Top10* can greatly reduce bias and is more efficient than conventional summary estimators. Recall that precision is the inverse of the estimate's standard error, or $1/SE_i$. We do not wish to supplant existing corrections for publication selection such as: trim and fill, funnel-asymmetry and precision-effect meta-regression analysis, Heckman meta-regression or maximum likelihood selection models (Duval and Tweedie, 2000; Stanley, 2005; Stanley, 2008; Moreno et al., 2009; Hedges, 1997). Rather, we use this paradox and its simple remedy to highlight the widespread vulnerability of the empirical sciences to publication selection. In the process, we underscore the critical importance of the oft neglected statistical dimension, precision, in improving scientific inference.

2. Funnel Graphs and Precision

> "The simplest and most commonly used method to detect publication bias is an informal examination of a funnel plot." – Sutton et al. (2000, p.1574)

2.1 Funnel Graphs and Publication Selection

Funnel graphs have been widely used in medical research and the social sciences to identify publication selection. A funnel graph is a scatter diagram of a reported empirical estimate (e_i) and its precision (*i.e.*, the inverse of the estimate's standard error, or $1/SE_i$). As its name suggests, a funnel plot should resemble an inverted funnel (see Figure 1). As the estimates become more precise (*i.e.*, moving from the bottom to the top of the diagram), the reported estimates become less spread out and tend to converge to the 'true' value. In the absence of publication selection (or selection for statistical significance), the plot will be symmetric—see Figure 1.

{Insert Figure 1 about there}

The idea that symmetry is implied by the absence of publication selection assumes, of course, that there is only one underlying population from which each estimate is drawn (*i.e.*, homogeneity). In many areas of empirical research, this assumption will not be valid, and multivariate meta-regression analysis will be required. Multivariate meta-regression is routinely employed in economics to explain the widely observed, systematic heterogeneity among reported empirical estimates. A funnel graph can also be used to identify when heterogeneity is required to be addressed explicitly. If the funnel graph has no single peak, no single parameter will adequately summarize this area of research, and heterogeneity will need to be explicitly modeled.

{Insert Figure 2 about there}

Publication selection for a specific directional effect (whether positive or negative) will skew the reported results and make the funnel graph asymmetric. Asymmetry is the hallmark of publication selection (Sutton et al., 2000), and it is routinely observed in the majority of areas of economic research (Stanley and Doucouliagos, 2008). See Figures 2, 3 and 4 for several examples. If the funnel graph is 'inverted' by placing SE_i on the vertical axis and then the axes are reversed, the funnel graph can be fitted by meta-regression analysis (MRA) (Card and Kreuger, 1995; Egger et al., 1997; Stanley, 2005; Stanley, 2008):

$$e_i = \beta_e + \beta_{SE} SE_i + \varepsilon_i \quad ; \quad i=1, 2, \ldots L \tag{1}$$

where i is an index that denotes a given study's reported estimate in a research literature comprised of L studies.

Equation (1) will contain obvious heteroscedasticty; thus, weighted least squares (WLS) are almost always employed when this MRA model is estimated. The WLS meta-regression model (1) may be expressed as a regression of a study's reported t-values (t_i) on precision ($1/SE_i$).

$$t_i = \beta_{SE} + \beta_e (1/SE_i) + \nu_i \tag{2}$$

(Egger et al., 1997; Stanley, 2008). Testing whether $\beta_{SE}=0$ provides an objective test for the funnel asymmetry and therefore for the presence of publication bias. Simulations show that this funnel-asymmetry test (FAT) is valid, although it has low power in identifying publication selection (Stanley, 2008).

Testing H_0: $\beta_e =0$ serves as a powerful test of whether there is a genuine empirical effect beyond publication selection (Stanley, 2008). Medical researchers use the estimate of β_e in (2) as a corrected empirical effect (Sutton et al., 2000; Moreno et al., 2009). However, this estimate is known to be biased downward when there is a genuine nonzero effect (Stanley, 2008).

{Insert Figure 3 about there}
{Insert Figure 4 about there}

The bias induced by publication selection often exceeds the magnitude of the underlying phenomenon being estimated (Stanley and Doucouliagos, 2008; Doucouliagos and Stanley, 2009). Typical funnel graphs for a given area of empirical research are skewed, often highly so, as Figures 2, 3 and 4 show. The real problem for empirical science is that conventional summary statistics can be greatly distorted. For example, the rates of smoking cessation from nicotine replacement therapy (NRT) are highly skewed (Figure 2), and the beneficial effect

of this therapy is greatly diminished once one accounts for publication selection (see the next section).

In the case of the employment effect of minimum-wage raises (Figure 3), the average reported elasticity is -0.19, which is statistically quite significant and widely regarded by economists to be an important adverse effect. Elasticity is the standard way economists measure the effect of one variable on another, controlling for all other factors. This average elasticity estimates the percent decrease in employment (about 0.2%) that would result from a one percent increase in the minimum wage. However, once publication selection is accounted for, little or no evidence of any adverse employment effect remains (Doucouliagos and Stanley, 2009).

For the effect of adopting a common currency (*e.g.*, the Euro) on the flow of trade, the average reported gamma, which is the regression coefficient from a logarithmic relation, is 0.859 and implies that a 136% increase in trade would result from joining a currency union—see Figure 4. This is almost universally regarded as a very large and practically important effect. Needless to say, it is also very statistically significant using conventional summary statistics. However, judging by the top of the graph (Figure 4), it is not at all clear that there is much, if any, trade effect from adopting a common currency.

2.2 On the Importance of Being Precise

Note how the tops of the funnel graphs are more tightly distributed and less skewed. These more precise estimates may still be biased, in the case of Figures 2, 3 and 4, but the bias will be of a much smaller magnitude. Not only are precise estimates more reliable and more efficient, they also have a smaller bias when the results are selected for their statistical significance. If there is a small empirical effect but the estimate's standard error is much smaller still, there will be virtually no need for selection and practically no bias. Even when the true effect is zero, the selection for the significance of very precise estimates will induce only a small bias. The point to these obvious observations is that precision is a key dimension of publication selection and its bias. Thus, the top of a funnel graph deserves special

attention and greater weight. Reviewers and journal editors should use precision as one objective measure of a study's quality.

Meta-analysts have long recognized the role that precision can play in summarizing an area of research and in handling publication selection bias, recall equation (2) (Egger et al., 1997; Stanley, 2008). The MRA test, H_0: β_e=0, for the presence of an empirical effect beyond publication bias has been called the precision-effect test (PET), because it is the MRA regression coefficient on precision (Stanley, 2008). When applied to NRT clinical trials, PET provides borderline evidence of a genuine positive effect of using the patch for smoking cessation. Uncorrected, the average risk ratio for NRT is 1.93; that is, the experimental group had a 93% higher smoking cessation rate than the control group, on average. However, PET is only marginally significant (t=2.00; p=.053). Although one might be tempted to use the fact that there are only 42 controlled studies that use the patch as NRT to explain this marginal result, this precision-effect test (H_0: β_e =0) has been found to be powerful even in smaller samples (Stanley, 2008). In contrast, there is no ambiguity about the presence of publication selection. The funnel-asymmetry test (H_0: β_{SE}=0) shows clear signs of selection bias (t=3.02; p<.01). Unlike the precision-effect test, this test (FAT) is known to have low power (Egger et al., 1997; Stanley, 2008) and yet there is clear evidence of publication selection. All three research areas displayed in funnel graphs 2, 3, and 4 contain clear evidence of asymmetry and hence publication selection.

Meta-analysts also exploit precision in the conventional summary weighted averages called 'fixed- and random-effect' estimators (FEE and REE, respectively). FEE weights each reported estimate by the inverse of the square of its standard error—or its precision squared. Weighting by the precision squared has been shown to be efficient (Cooper and Hedges, 1994). FEE assumes that all of the reported estimates are drawn from the same population with a common mean. When estimates are drawn from several populations (*i.e.*, when there is heterogeneity), REE becomes the appropriate estimator. It weights each estimate by the inverse of a more complex variance that contains two components: $SE_i^2 + S_h^2$; where S_h^2 is an estimate of the between-study or heterogeneity variance.

Because both of these summary statistics give greater weight to more precise estimates, they are less biased than the simple mean when there is publication selection. Clearly, precision is the key to a less biased and more efficient summary of empirical research when results are selected, in part, for statistical significance.

2.3 Top10

What if the most precise estimates were given even greater weight? Wouldn't this reduce the effects of publication bias further? Taken to the extreme, we might give the most precise (say the most precise ten percent of the reported estimates) a weight of one and the remainder a weight of zero. Obviously, this is a radical proposal, one that goes against much statistical theory and practice. However, this *Top10* estimator can be shown to be less biased and more efficient than conventional summary estimators under conditions found in many areas of scientific research. We do not wish to imply that this naïve approach contains the most efficient set of weights for the reported empirical findings. The most efficient estimator would depend, among other things, on the proportion of scientific findings that are selected for their statistical significance. Unfortunately, this incidence of publication selection is inherently unobservable; thus, identifying the most efficient set of weights would be a difficult problem that serves little practical purpose. We chose the top 10% because it is less biased than larger percents (*e.g.*, the top 20%). However, as the chosen percentage decreases, the mean of the remaining few will be less and less efficient. The choice of the top percentage will remain somewhat arbitrary because its efficacy will ultimately depend, among other things, on the actual incidence of publication selection.

3. Simulations

The design used in these simulations is the same as that employed in Stanley (2008) to investigate the small-sample properties of testing the MRA parameters in equation (2) against zero. Random data are generated, and a regression coefficient is estimated and tested against zero (*i.e.*, H_0: β=0). Heterogeneity variation and

regression residuals are drawn from independent, normal distributions. Regression is chosen because it is a dominant statistical technique used in several social sciences and may be regarded as encompassing other widely applied techniques (*e.g*, ANOVA and t-tests) (Moore, 1997).

Publication bias is simulated as selecting a statistically significant positive regression coefficient. That is, if a random estimate does not provide a significantly positive t-value, a new sample is taken and the original regression is run again with different random errors and heterogeneity until a significant t-value is obtained by chance. For example, the 50% publication selection condition assumes that exactly half of the studies estimate and re-estimate their regression models until a random, yet significantly positive, estimate is found and reported. For the other half, the first random estimate, significant or not, is reported and used. In practice, not all published results will have been selected for statistical significance. Therefore, it is assumed that the incidence of publication selection is either: 0%, 25%, 50%, 75% or 100%.

The sample sizes used are 40 and 80. Many areas of economics research report more estimates, often many times more. The smaller sample size is chosen to be more consistent with medical research. In particular, there are 42 clinical trials of the effect on smoking cessation from nicotine replacement therapy (NRT) using the patch (n=42). When all nicotine delivery methods are combined, there are more than 100 clinical trials (n=112) (Stead et al., 2008). Changing the sample size has little effect on the bias or the relative performance of these estimators. Of course, reducing the sample size will increase the mean square errors of all estimators, especially the *Top10*. Throwing out 90% of the data is especially imprudent when the number of estimates available in a research literature is small.

Along with the incidence of publication selection, publication bias is most highly influenced by the magnitude of the unexplained heterogeneity relative to sampling error (σ_ε^2) (Stanley, 2008; Moreno et al., 2009). $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$ is a widely employed indicator of the magnitude of heterogeneity (Higgins and Thompson, 2002). It may be interpreted as the proportion of the observed variance ($\sigma_h^2+\sigma_\varepsilon^2$) that is due to heterogeneity across studies (σ_h^2). Tables 1 and 2 report this

heterogeneity proportion, I^2, as well as the incidence of publication selection for a wide variety of Monte Carlo experiments. I^2 is controlled by changing the between-study variance. It is calculated from the population parameters when there is no publication selection. Sample estimates of I^2 will vary as a function of selection and the existence of a true empirical effect.

Table 1: Means of Alternative Research Summary Estimators (n=80)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	$\hat{\beta}_e$
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.23	0.20	0.22	0.13	0.04
	0	50%	0.47	0.39	0.43	0.28	0.06
	0	75%	0.70	0.59	0.63	0.41	0.07
I^2=25%	0	100%	0.93	0.78	0.78	0.55	0.07
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.07	1.04	1.04	1.00	0.92
	1	50%	1.13	1.08	1.09	1.01	0.85
	1	75%	1.20	1.11	1.13	1.02	0.77
	1	100%	1.26	1.15	1.16	1.02	0.68
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.27	0.23	0.25	0.14	0.04
	0	50%	0.54	0.45	0.51	0.30	0.08
	0	75%	0.81	0.68	0.75	0.47	0.14
I^2=58%	0	100%	1.08	0.91	0.92	0.66	0.20
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.10	1.07	1.08	1.02	0.94
	1	50%	1.19	1.13	1.16	1.04	0.88
	1	75%	1.29	1.19	1.23	1.07	0.81
	1	100%	1.39	1.26	1.30	1.08	0.74
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.36	0.29	0.34	0.18	0.04
	0	50%	0.72	0.58	0.68	0.38	0.10
	0	75%	1.09	0.88	1.02	0.62	0.22
I^2=85%	0	100%	1.45	1.20	1.29	0.88	0.37
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.18	1.13	1.17	1.06	0.97
	1	50%	1.36	1.27	1.33	1.12	0.93
	1	75%	1.54	1.39	1.49	1.17	0.87
	1	100%	1.73	1.52	1.63	1.22	0.80

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. $\hat{\beta}_e$ is estimated from equation 2.

Tables 1 and 2 report the average values of alternative estimators observed over 10,000 replications. The true value of the underlying effect is either 0 or 1, as displayed in the tables. Note how all of these conventional summary statistics, the simple mean and various weighted averages (including the *Top10*), are biased

upward when there is some publication selection (selection incidence ≥25%). As expected, these biases will increase with the incidence of publication selection and the between-study heterogeneity.

Table 2: Means of Alternative Research Summary Estimators (n=40)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	$\hat{\beta}_e$
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.23	0.20	0.22	0.14	0.04
	0	50%	0.47	0.39	0.43	0.28	0.06
	0	75%	0.70	0.59	0.63	0.41	0.07
I^2=25%	0	100%	0.94	0.78	0.78	0.55	0.07
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.07	1.04	1.04	1.00	0.92
	1	50%	1.13	1.08	1.09	1.01	0.85
	1	75%	1.20	1.11	1.13	1.02	0.77
	1	100%	1.26	1.15	1.16	1.03	0.69
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.27	0.22	0.25	0.14	0.04
	0	50%	0.54	0.45	0.51	0.30	0.09
	0	75%	0.81	0.68	0.75	0.48	0.14
I^2=58%	0	100%	1.08	0.91	0.92	0.66	0.20
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.10	1.07	1.08	1.02	0.95
	1	50%	1.19	1.13	1.16	1.04	0.88
	1	75%	1.29	1.19	1.23	1.06	0.81
	1	100%	1.39	1.26	1.30	1.08	0.74
	0	0%	0.00	0.00	0.00	0.00	0.00
	0	25%	0.36	0.29	0.34	0.18	0.03
	0	50%	0.72	0.58	0.68	0.39	0.11
	0	75%	1.09	0.89	1.02	0.62	0.22
I^2=85%	0	100%	1.45	1.20	1.29	0.89	0.38
	1	0%	1.00	1.00	1.00	1.00	1.00
	1	25%	1.18	1.13	1.17	1.06	0.97
	1	50%	1.36	1.26	1.33	1.13	0.94
	1	75%	1.55	1.39	1.49	1.18	0.88
	1	100%	1.73	1.52	1.63	1.23	0.81

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. $\hat{\beta}_e$ is estimated from equation 2.

Whenever there is publication selection, the *Top10* has smaller bias than the conventional summary statistics, the simple mean and other weighted averages (FEE and REE), that use all the data. In many cases, *Top10* lowers the bias substantially even when compared to the weighted averages which give greater weight to the more precise estimates (FEE and REE). When there is no publication selection, all estimators are virtually unbiased. Thus, Tables 1 and 2 shows that

there is something to be said for discarding 90% of research when there is selection for statistical significance.

Table 3: Mean Square Errors of Alternative Research Summary Estimators (times 1,000 with n=80)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	$\hat{\beta}_e$
	0	0%	3	3	3	14	27
	0	25%	58	41	49	33	25
	0	50%	221	155	186	88	22
	0	75%	494	344	396	180	17
I^2=25%	0	100%	875	603	603	310	10
	1	0%	3	3	3	14	27
	1	25%	7	4	5	13	30
	1	50%	20	8	10	13	45
	1	75%	41	15	18	13	74
	1	100%	71	25	28	13	115
	0	0%	6	6	6	27	51
	0	25%	78	55	69	49	45
	0	50%	295	207	260	115	43
	0	75%	658	464	560	243	44
I^2=58%	0	100%	1168	830	839	447	51
	1	0%	6	6	5	27	50
	1	25%	15	9	11	26	50
	1	50%	42	22	30	25	56
	1	75%	88	42	58	25	73
	1	100%	152	70	92	26	99
	0	0%	16	14	14	64	116
	0	25%	145	94	129	97	104
	0	50%	535	344	477	206	94
	0	75%	1087	788	1040	422	108
I^2=85%	0	100%	2100	1436	1674	792	172
	1	0%	16	14	14	63	114
	1	25%	46	30	39	59	101
	1	50%	144	80	120	63	93
	1	75%	305	161	245	71	88
	1	100%	534	273	406	82	95

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. $\hat{\beta}_e$ is estimated from equation 2.

Bias is one thing, but efficiency can be quite another. Surely, discarding 90% of the data cannot be efficient. Tables 3 and 4 report the mean squared errors (MSE) of these conventional summary statistics. Surprisingly, the *Top10* is also more efficient, as defined by smaller MSE, in many cases. Essentially, as long as 50% or more of the studies use statistical significance as one criterion for reporting results, the *Top10* will have lower MSE than conventional summary statistics. This is certainly the case when compared to the simple average, but FEE and/or REE may

have slightly smaller MSE than *Top10* when the incidence of publication selection is exactly 50% and there is a genuine effect. The relative performance of *Top10* depends on the amount of heterogeneity. As heterogeneity increases, expected publication bias worsens; thus, the relative performance of the *Top10* improves. In over half the cases reported in Tables 3 and 4, the *Top10* has the smallest MSE among all these averages, simple and weighted. However, we are not trying to prove that *Top10* is the best estimator, only that discarding 90% of the data may be a feasible strategy in some cases.

Table 4: Mean Square Errors of Alternative Research Summary Estimators (times 1,000 with n=40)

Hetero-geneity*	True effect	Selection Incidence	Simple Average	FEE	REE	*Top10*	$\hat{\beta}_e$
	0	0%	7	6	4	28	54
	0	25%	60	43	51	50	48
	0	50%	223	157	188	102	40
	0	75%	492	343	395	190	30
I^2=25%	0	100%	876	604	605	314	15
	1	0%	7	6	6	28	55
	1	25%	10	7	7	27	55
	1	50%	22	10	12	26	70
	1	75%	44	18	21	25	96
	1	100%	72	27	30	24	137
	0	0%	12	11	10	54	104
	0	25%	83	60	73	79	91
	0	50%	299	211	263	144	80
	0	75%	662	467	562	265	69
I^2=58%	0	100%	1168	831	846	456	63
	1	0%	11	11	10	55	103
	1	25%	20	14	16	51	100
	1	50%	46	26	34	48	98
	1	75%	91	47	61	47	107
	1	100%	155	74	96	47	129
	0	0%	32	29	28	131	234
	0	25%	160	107	142	168	211
	0	50%	543	354	483	274	184
	0	75%	1195	797	1047	479	177
I^2=85%	0	100%	2104	1444	1678	829	209
	1	0%	31	28	27	132	238
	1	25%	61	43	52	123	210
	1	50%	179	91	129	118	179
	1	75%	316	173	255	123	156
	1	100%	542	282	414	123	149

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. $\hat{\beta}_e$ is estimated from equation 2.

Thus, we have demonstrated the genuine *possibility* that discarding 90% of the data can improve our scientific knowledge. Although it is impossible to know the true prevalence of publication selection in the empirical sciences, we suspect that many specific areas of research will have the majority of their results subject to publication selection. Among clinical medical trials, "trials with positive findings . . . had nearly four times the odds of being published compared to findings that were not statistically significant" (Hopewell *et al*., 2009, Summary). In two-thirds of the areas of economics research, we find either 'substantial' or 'severe' publication selection (Stanley and Doucouliagos, 2008). Such high levels of selection, as measured by the estimate of β_{SE} from equation (2), are likely to correspond to an incidence of selection of 50% or greater (Stanley and Doucouliagos, 2008, Appendix Table 1).

Thus far, we have compared the *Top10* to simple and weighted averages of reported effects. But none of these alternative estimators are specifically designed to correct for publication bias. It remains to be seen how the *Top10* performs relative to meta-analytic methods designed to reduce publication bias. Economists and medical researchers use estimates of the MRA coefficient, β_e, from equation (2) to correct for publication bias (Sutton et al., 2000; Stanley, 2008; Moreno et al., 2009). A 'comprehensive simulation study' by medical researchers concludes that: "Several of the regression based methods consistently outperformed the Trim & Fill estimators" (Moreno et al., 2009, Results). Of these regression-based corrections of publication bias, the coefficient on precision in MRA equation (2), β_e, has been used most often. The last column of Tables 1-4 report means and MSE of the estimates of this MRA coefficient on precision, $\hat{\beta}_e$. Whenever there is a genuine empirical effect, *Top10* has a lower MSE than does $\hat{\beta}_e$. However, when there is no genuine underlying empirical effect, this relative performance mostly reverses (see Tables 1-4). In over half of these cases (39 out of 60), *Top10* has a lower MSE than a MRA estimator that is designed to reduce publication bias and uses ***all*** of the reported results.

The practical utility of employing the *Top10* is considerably greater that this direct comparison to $\hat{\beta}_e$ would seem to indicate. We only need a corrected estimate of the overall empirical effect when we have reason to believe that there is, in fact, some nonzero empirical effect. As these simulations show, *Top10* will be less biased than $\hat{\beta}_e$ in the majority of these cases. Previous simulations have demonstrated that testing H_0: β_e=0 (*i.e.*, precision-effect test) often serves as a powerful and valid test of the presence of a genuine empirical effect beyond publication selection (Stanley, 2008). Thus, in most of those cases where we reject H_0: β_e=0, *Top10* will be less biased. In the absence of evidence of an authentic effect (*i.e.*, the failure to reject H_0: β_e =0), all estimators will be biased, and we are better off assuming that β_e is zero (Stanley, 2008).

Lastly, Table 5 reports coverage probabilities, which measure the proportion of the simulations (replications=10,000) where the true effect falls with the calculated 95% confidence interval. In addition to *Top10* and $\hat{\beta}_e$, the random-effects estimator (REE) is included. When there is excess random heterogeneity, as is the case for all of these simulations, REE (rather that FEE or the simple mean) is the valid summary statistic. Therefore, REE is likely to possess better coverage properties than either the simple average or FEE. As before, when there is a genuine empirical effect, *Top10* performs the best. In this critical situation, *Top10* has excellent coverage probabilities even in the presence of dominating publication selection. However, in these same cases, the coverage for both $\hat{\beta}_e$ and REE is often unacceptably low.

When there is no true empirical effect, the tables are turned, and *Top10* has unacceptably low coverage probabilities in most cases. Even here, *Top10* consistently performs better than the random-effects estimator (REE). In the majority of all the cases simulated, *Top10*'s coverage is closer to the nominal level (95%) than is $\hat{\beta}_e$. Nonetheless, *Top10*'s low coverage probabilities in some cases would constitute a serious cause of concern if *Top10* were to be used as a corrected estimate of overall effect when there is no actual underlying empirical effect. But a

corrected estimate is only needed when we have evidence that indicates the existence of a true empirical effect (reject H_0: β_e=0).

Table 5: Coverage of Alternative Research Summary Estimators (n=80)

Hetero-geneity*	True effect	Selection Incidence	REE	*Top10*	$\hat{\beta}_e$
	0	0%	.950	.949	.936
	0	25%	.038	.879	.984
	0	50%	0	.582	.988
	0	75%	0	.179	.985
I^2=25%	0	100%	0	0	.815
	1	0%	.947	.955	.939
	1	25%	.863	.954	.907
	1	50%	.580	.949	.811
	1	75%	.221	.953	.607
	1	100%	.042	.944	.307
	0	0%	.954	.951	.934
	0	25%	.081	.900	.971
	0	50%	0	.676	.969
	0	75%	0	.276	.941
I^2=58%	0	100%	0	0	.489
	1	0%	.950	.948	.928
	1	25%	.793	.950	.929
	1	50%	.322	.949	.886
	1	75%	.024	.941	.791
	1	100%	0	.934	.590
	0	0%	.952	.939	.932
	0	25%	.170	.909	.959
	0	50%	0	.732	.957
	0	75%	0	.367	.895
I^2=85%	0	100%	0	0	.333
	1	0%	.937	.954	.933
	1	25%	.673	.952	.940
	1	50%	.099	.936	.936
	1	75%	0	.915	.902
	1	100%	0	.891	.790

* Heterogeneity is measured by $I^2=\sigma_h^2/(\sigma_h^2+\sigma_\varepsilon^2)$. FEE & REE denote the fixed-effects and random-effects estimators, respectively. $\hat{\beta}_e$ is estimated from equation 2.

Ignoring publication selection can result in large biases and in the misidentification of empirical effects that do not actually exist. Such biases can lead to inappropriate policy decisions even when based on all applicable research. Yet doing something as simple as calculating the average of the most precise 10% of the reported estimates can greatly reduce this bias and improve policy. If nothing else, policy makers and practitioners could use the difference between the *Top10*

and the simple average as an indicator of the presence of publication biases. When this difference is of a practically important magnitude, the *Top10* or a more sophisticated publication bias correction technique should be employed. There are other, more sophisticated, meta-regression methods that will outperform both *Top10* and $\hat{\beta}_e$ in some cases, but no single method is dominant (Moreno et al., 2009; Stanley and Doucouliagos, 2007).

4. What's the Difference? Policy Implications of Publication Selection

The difference between these estimates often has practical consequences. In 2007, the US Congress raised the minimum wage, starting July 24th 2007, by $.70 per year for three years, until it reaches a value of $7.25/hour in 2009. Each time that new minimum-wage legislation is brought before the US Congress, opposition cites economics research and the consensus among economists that raising the minimum wage will cause a decrease in employment (or an increase in unemployment). Economists' reasoning is very simple and based on supply and demand. Essentially, if you raise the cost of hiring (and keeping) workers, businesses will be able to afford fewer of them. A discussion of the adverse employment effects of raising the minimum wage is contained in virtually every introductory economics textbook.

In spite of 678 reported statistically significant estimates of minimum-wage's adverse employment effect, a more comprehensive analysis of the reported research finds "no evidence of a meaningful adverse employment effect when selection effects are filtered from the research record" (Doucouliagos and Stanley, 2009, pp. 422-23). For example, the *Top10* estimate of this minimum-wage effect is -0.0217 (implying that a doubling of the minimum wage will lead to a 2% decrease in teenage employment). When a multivariate meta-regression approach is used that explicitly models both publication selection and the underlying empirical effect, adverse employment effects are converted to a *positive* and significant employment effect (Doucouliagos and Stanley, 2009). There are theoretical reasons for this *positive* effect on employment. For example, workers might become more loyal

and more productive when they are paid well, the 'efficiency-wage hypothesis.' Furthermore, there is empirical support for such a *positive* effect of wage raises among tests of the efficiency–wage hypothesis after correcting for publication selection (Krassoi Peach and Stanley, 2009). The point is that a closer examination of the 1,474 reported estimates of minimum-wage employment effects provides little justification for meaningful adverse minimum wage effect. The simple *Top10* reduces the minimum-wage effect to practically zero.

Likewise, the *Top10* greatly reduces the practical consequences of the adoption of a common currency. Recall that the average reported common currency effect is 0.859 (or 136%). If their trade with the EU were really to increase by 136%, the UK and Denmark (for example) would find it extremely difficult to continue to resist the pull of the EMU (European Monetary Union) and the Euro. Such a large economic benefit might easily be sufficient to persuade policy makers in the UK and Denmark to forgo much of their economic policy independence. In contrast, the *Top10* estimates the trade effect attributable to joining a currency union to be only 10%. This smaller economic benefit of currency union may be seen by some policy makers as inadequate compensation for the accompanying reduction of national economic sovereignty. Worse still, testing H_0: β_e=0 in equation (2) provides evidence of a *negative* trade effect (t=-4.36; p<.01) after correcting for publication selection. Recall that the above simulations show that that the *Top10* is biased upwards, when there is no underlying empirical effect. Some upward or positive bias could remain even if the true empirical effect were slightly negative. Correcting for publication selection with even this simple and naïve estimator can have important policy consequences.

Lastly, correcting for publication selection also has a large practical effect on the efficacy of using a nicotine replacement patch for smoking cessation (recall Section 2.2). As discussed above, the unadjusted average risk ratio is 1.93 and is only marginally statistically significant when allowance is made for publication selection. The *Top 10* reduces this risk ratio to 1.53, which nearly halves the improved likelihood of quitting smoking. Although nicotine replacement therapy

still shows some efficacy, its advantage over alternative approaches becomes much less clear.

5. Conclusions

Could it be better to discard 90% of the reported research? Surprisingly, the answer is yes to this statistical paradox. This paper has shown how publication selection can greatly distort the research record and its conventional summary statistics. Using both Monte Carlo simulations and actual research examples, we show how a simple estimator, which uses only 10 percent of the reported research reduces publication bias and improves efficiency over conventional summary statistics that use *all* the reported research.

The average of the most precise 10 percent, '*Top10,*' of the reported estimates of a given empirical phenomenon is often better than conventional summary estimators because of its heavy reliance on the reported estimate's precision (*i.e.*, the inverse of the estimate's standard error). When estimates are chosen, in part, for their statistical significance, studies cursed with imprecise estimates have to engage in more intense selection from among alternative statistical techniques, models, data sets, and measures to produce the larger estimate that statistical significance demands. Thus, imprecise estimates will contain larger biases.

Studies that have access to more data will tend to be more precise, and hence less biased. At the level of the original empirical research, the statistician's motto, "the more data the better," holds because more data typically produce more precise estimates. It is only at the meta-level of integrating, summarizing, and interpreting an entire area of empirical research (meta-analysis), where the removal of 90% of the data might actually improve our empirical knowledge. Even when the authors of these larger and more precise studies actively select for statistical significance in the desired direction, smaller significant estimates will tend to be reported. Thus, precise studies will, on average, be less biased and thereby possess greater scientific quality, *ceteris paribus*.

We hope that the statistical paradox identified in this paper refocuses the empirical sciences upon precision. Precision should be universally adopted as one criterion of research quality, regardless of other statistical outcomes.

References:

American Psychological Association. (1994), *Publication Manual of the American Psychological Association* (4th ed.). Washington.

Begg, C. B. and Berlin, J.A. (1988), "Publication Bias: A Problem in Interpreting Medical Data," *Journal of the Royal Statistical Society A*, 151, 419-445.

Card, D. and Krueger, A.B. (1995), "Time-Series Minimum-Wage Studies: A Meta-Analysis," *American Economic Review*, 85, 238-243.

Carver, R.P. (1978), "The Case Against Statistical Testing," *Harvard Educational Review*, 48, 378-399.

Cohen, J. (1994), "The Earth Is Round ($p < .05$)," *American Psychologist*, 49, 997-1003.

Cooper, H.M. and Hedges, L. V. (eds.) (1994), *Handbook of Research Synthesis.* New York: Russell Sage

Copas, J. (1999), "What Works? Selectivity Models and Meta-Analysis," *Journal of the Royal Statistical Society*, A , 161, 95-105.

Costa-Font, J., Gammill, M. and Rubert, G. (2008), "Re-Visiting the Health Care Luxury Good Hypothesis: Aggregation, Precision and Publication Bias," Doucments de Treball, E08/197.

Daniel, L.G. (1998), "The Statistical Significance Controversy Is Definitely Not Over: A Rejoinder to Responses by Thompson, Knapp, and Levin," *Research in the Schools*, 5, 63-65.

De Long, J.B. and Lang, K. (1992), "Are All Economic Hypotheses False?" *Journal of Political Economy*, 100, 1257-72.

Doucouliagos, C. (H) and Laroche, P. (2003), "What Do Unions Do To Productivity: A Meta-Analysis, *Industrial Relations*, 42, 650-691.

Doucouliagos, C. (H), Laroche, P. and Stanley T.D. (2005), "Publication Bias in Union-Productivity Research," *Relations Industrielles/Industrial Relations*, 60, 320-346.

Doucouliagos, C. (H) and Stanley, T.D. (2008), Theory Competition and Selectivity. Deakin Working Paper, Economics Series 2008-06.

Doucouliagos, C.(H) and Stanley, T.D. (2009), "Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis," *British Journal of Industrial Relations*, 47, 406-29.

Duval, S. and R. Tweedie (2000), "A Nonparametric Trim and Fit Method of Accounting for Publication Bias in Meta-Analysis," *Journal of the American Statistical Association*, 95, 89-98.

Egger, M., Smith, G.D., Scheider, M., and Minder, C. (1997), "Bias in Meta-Analysis Detected by a Simple, Graphical Test," *British Medical Journal*, 316, 629-34.

Feige, E.L. (1975), "The Consequence of Journal Editorial Policies and a Suggestion for Revision," *Journal of Political Economy*, 83, 1291-5.

Gerber, A.S., Green, D.P. and Nickerson, D. (2001), "Testing for Publication Bias in Political Science," *Political Analysis*, 9, 385-92.

Gerber, A.S. and Malhorta, N. (2008), "Publication Bias in Empirical Sociological Research," *Sociological Methods & Research*, 37, 3-30.

Glass, G.V., McGaw, B. and Smith, M.L. (1981), *Meta-Analysis in Social Research*. Beverly Hills: Sage.

Harlow, L. L., Mulaik, S. A., and Steiger, J. H. (Eds.) (1997), *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum.

Hedges, L.V. (1992), "Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science*, 7, 246-55.

Hedges, L. V. and Olkin, I. (1985), *Statistical Methods for Meta-Analysis.* Orlando: Academic Press.

Higgins J.P.T. and Thompson S.G. (2002), "Quantifying Heterogeneity in Meta-Analysis. *Statistics in Medicine*, 21, 1539-1558.

Hopewell, S., Loudon, K., Clarke, M.J., Oxman, A.D. and K. Dickersin, K (2009), "Publication Bias in Clinical Trials due to Statistical Significance or Direction of Trial Result," *Cochrane Review*, Issue 1.http://www.thecochranelibrary.com

Krakovsky, M. (2004), "Register or Perish," *Scientific American*, 291(Dec.), 18-20.

Krassoi Peach, E. and Stanley, T.D. (2009), "Efficiency Wages, Productivity and Simultaneity: A Meta-Regression Analysis," *Journal of Labor Research*, forthcoming.

Light, R.J. and Pillemer, D.B. (1984), *Summing Up: The Science of Reviewing Research*. Cambridge, Mass.: Harvard University Press.

Lovell, M.C. (1983), "Data Mining," *The Review of Economics and Statistics,* 65, 1-12.

Moreno, S.G., Sutton, A.J., Ades, A., Stanley, T.D Abrams, K.R., Peters, J.L., and Cooper, N.J. (2009), "Assessment of Regression-Based Methods to Adjust for Publication Bias through a Comprehensive Simulation Study," *BMC Medical Research Methodology*, 9:2, http://www.biomedcentral.com/1471-2288/9/2.

Moore, D.S. (1997), "Bayes for Beginners? Some Reasons to Hesitate," *The American Statistician,* 51, 254-61.

Roberts, C. J. and Stanley, T.D. (eds) (2005), *Meta-Regression Analysis: Issues of Publication Bias in Economics*. Oxford: Blackwell.

Rose, A.K. and Stanley, T.D. (2005), "A Meta-Analysis of the Effect of Common Currencies on International Trade," *Journal of Economic Surveys,* 19, 347-65.

Rosenthal, R. (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86, 638-41.

Stanley, T.D. (2001), Wheat from Chaff: Meta-Analysis as Quantitative Literature Review," *Journal of Economic Perspectives*, 15, 131-50.

Stanley, T.D. (2005), "Beyond Publication Selection," *Journal of Economic Surveys*, 19, 309-345.

Stanley, T.D. (2008) "Meta-Regression Methods for Detecting and Estimating Empirical Effect in the Presence of Publication Bias. *Oxford Bulletin of Economics and Statistics*, 70,103-127.

Stanley T.D. and C. Doucouliagos (2007), "Identifying and Correcting Publication Selection Bias in the Efficiency-Wage Literature: Heckman Meta-Regression**,"** School Working Paper, Economics Series 2007-11, Deakin University.

Stead, L.F., Perera, R., Bullen, C., Mant, D., and Lancaster, T. (2008), "Nicotine Replacement Therapy for Smoking Cessation," *The Cochrane Library*, Issue 2, http://www.thecochranelibrary.com.

Sterling T.D. (1959), "Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance," *Journal of the American Statistical Association, 54*, 30-34.

Sterling, T.D., Rosenbaum, W.L. and Weinkam, J.J. (1995), "Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *American Statistician*, 439, 108-112.

Sutton, A.J., Abrams, K.R., Jones, D. R., Sheldon, T. A. and Song, F. (2000), *Methods for Meta-analysis in Medical Research,* Chichester: John Wiley and Sons.

Tullock, G. (1959), "Publication Decisions and Tests of Significance – A Comment, "*Journal of the American Statistical Association* 54, 593.

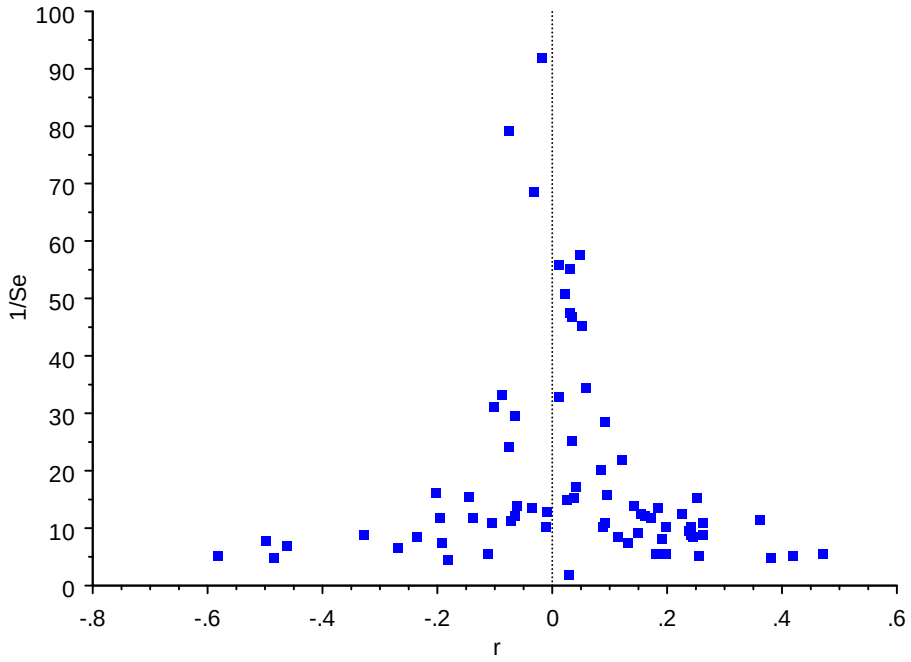

Figure 1. Funnel Plot of Union-Productivity Partial Correlations. A funnel plot is a scatter diagram of precision, the inverse of an estimate's standard error *vs.* its magnitude. Source: Doucouliagos and Laroche (2003)

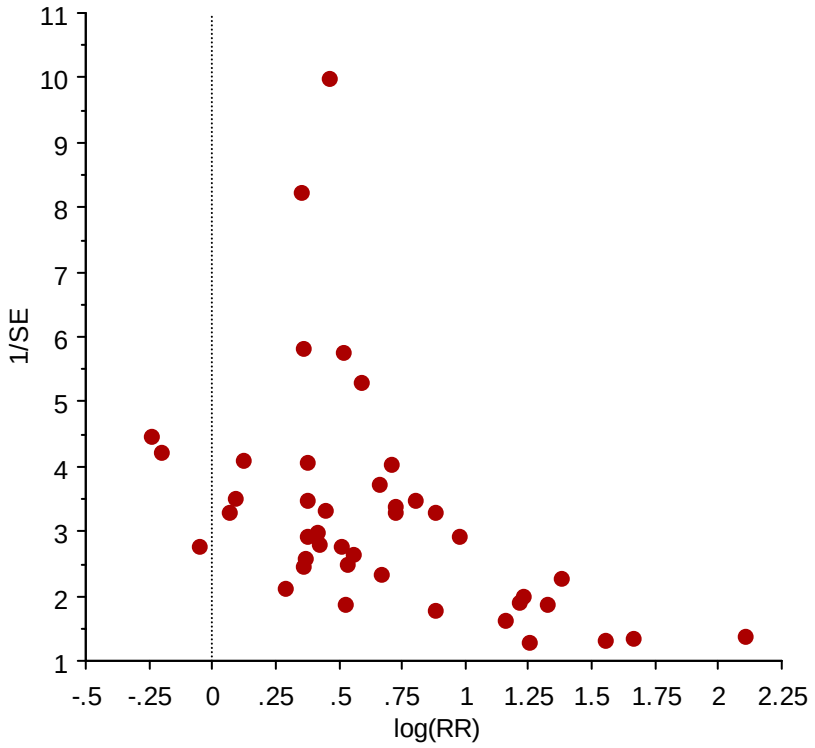

Figure 2: Funnel Graph of the Log Risk Ratio of Nicotine Replacement Therapy Using a Patch for Smoking Cessation (n=42). Source: Stead et al. (2008).

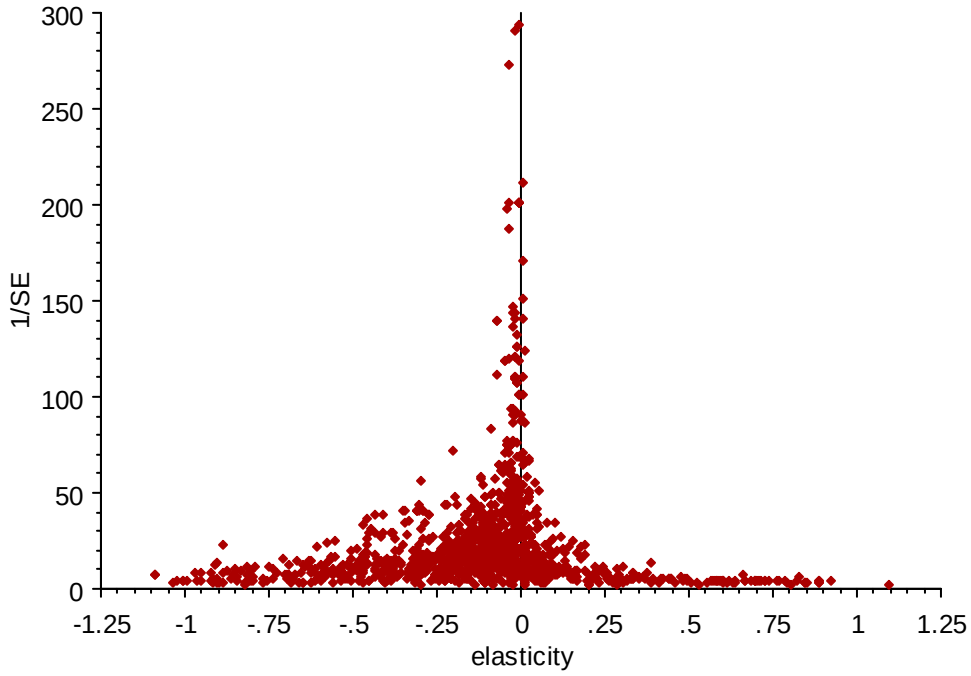

Figure 3: Trimmed Funnel Graph of Estimated Minimum-Wage Effects (n=1,424). Source: Doucouliagos and Stanley (2009). A few (50 or 3.4%) of the most extreme wage elasticities have been trimmed to reveal how the majority of this research is distributed. If the data were not trimmed, the graph would appear as a large spike in the middle with a handful of points on the extreme 'wings.'

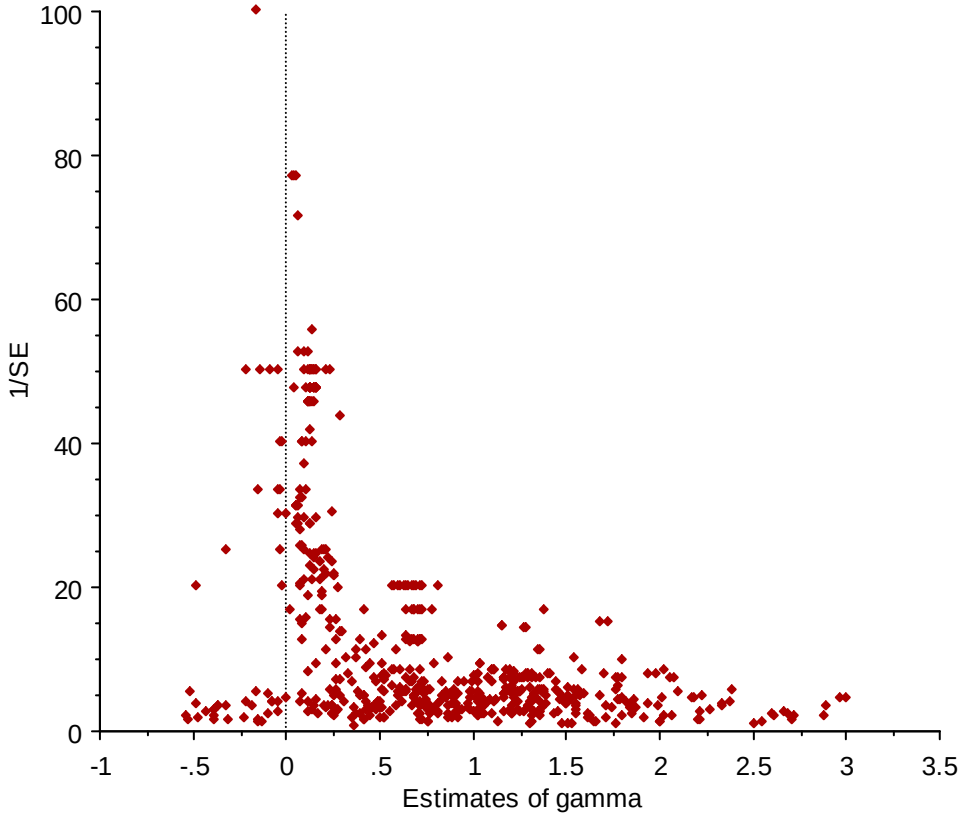

Figure 4. Trimmed Funnel Graph of Common Currency–Trade Effect. Source: Rose and Stanley (2005). A few (5%) extreme negative estimates as well as a few (5%) positive values have been trimmed in order to see the shape of the vast majority of these estimates of the trade effect.